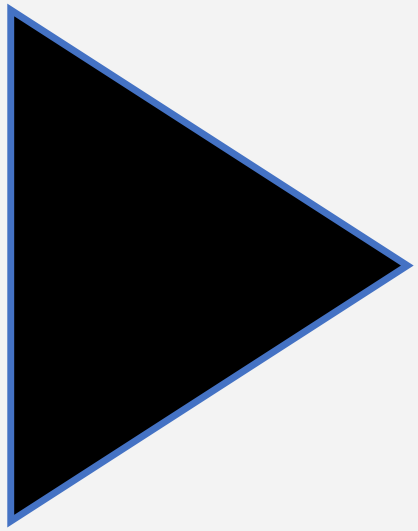
AI Factory Austria AI:AT

Building Knowledge from Text

Methods, Models, and Practice

Methods, Models, and Practice

Building Knowledge from Text



The goal is to turn data into information, and information into insight.

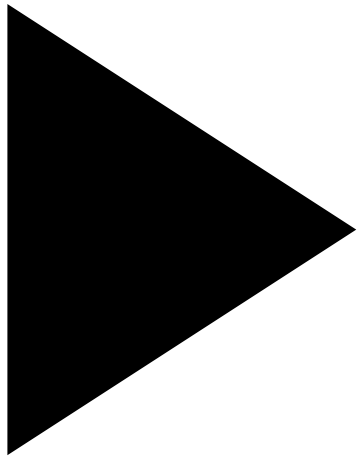
-- Carly Fiorina

We are overwhelmed by knowledge, e.g., articles, reports, notes, guidelines, websites

Yet:

- It takes too long to read
- It is hard to analyze systematically
- It often remains unused

How can we make large-scale knowledge accessible and usable?



Unstructured vs Structured

Unstructured

- Text-heavy
- Heterogeneous
- Human-readable

Structured

- Tabular
- Standardized
- Machine-readable

Key Concepts

Entity

An entity is a real-world object or concept (e.g. Person, Org. Location, Date)

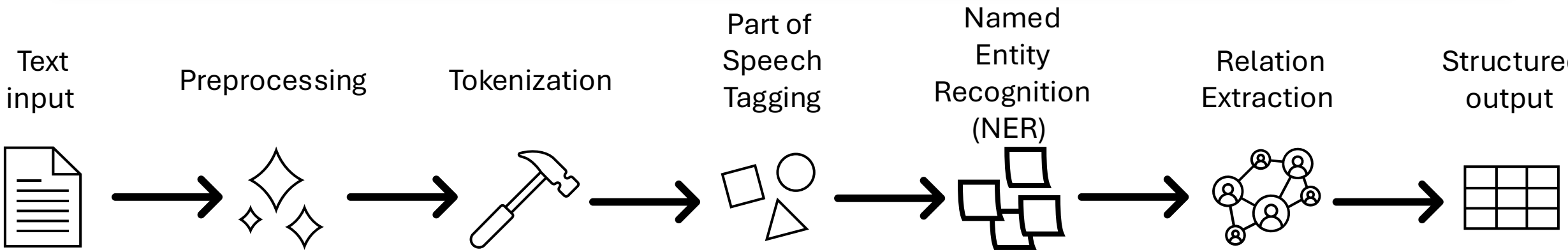
Relation

A relation describes an association between entities (e.g. works_for, located_in)

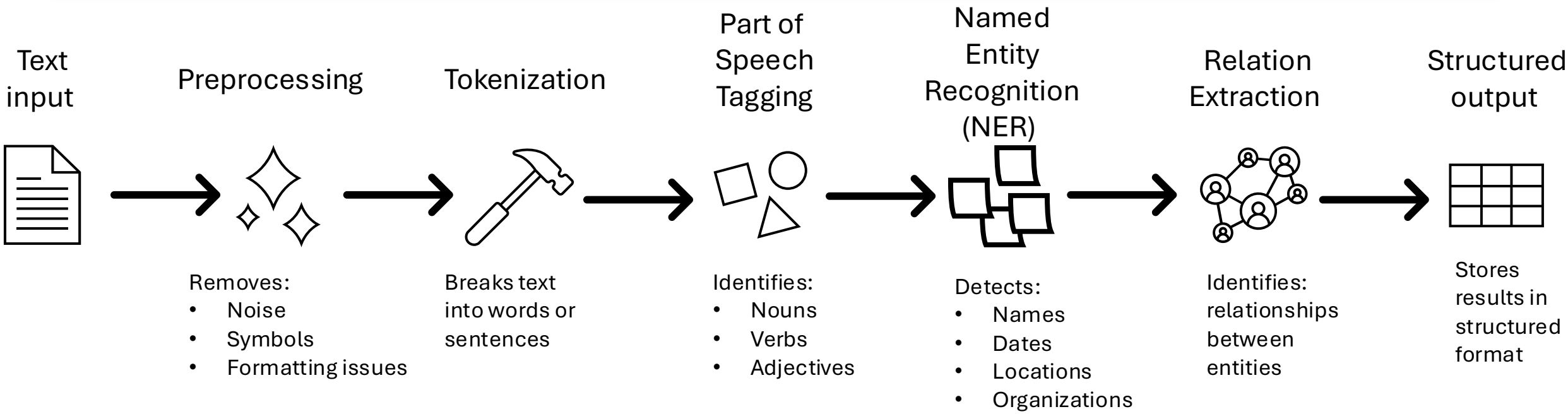
Fact

A fact is a structured statement usually represented as (Head Entity, Relation, Tail Entity)

Information Extraction Pipeline



Information Extraction Pipeline (cont.)



Generate Schema

- Subject matter experts
- Identify what information is needed

- Paragraphs / Words
- Keywords
- Topics

Applications

- Information Retrieval
- Search Engines/ Knowledge Graphs
- Biomedical and Healthcare
- Finance and Regulatory Compliance
- Legal and Scientific Research

Limitations

- Linguistic Ambiguity
- Data Quality and Availability
- Model Generalization and Scalability
- Error Propagation
- Domain Dependency

Methods, Models, and Practice

Simple pattern-based methods

- Regex
- Rules

Simple pattern-based methods

Simple pattern-based methods rely on predefined rules or patterns to identify and extract specific pieces of information from unstructured or semi-structured text

Advantages

- Transparency
- Control
- ability to handle specific domain knowledge
- no training data required
- rapid development and prototyping

Disadvantages

- scalability and adaptability
- maintenance overhead
- Brittleness
- limited generalization
- knowledge acquisition bottleneck

Regular expressions (Regex)

- Formal language for describing patterns in strings
- Uses metacharacters to represent classes of characters, repetition, position and grouping
- Metacharacters: ^ $ * + ? { } [ ] \ | ( )
- Limitations: readability / maintainability, performance issues, theoretical limitations, overuse and misapplication

Regex (cont.)

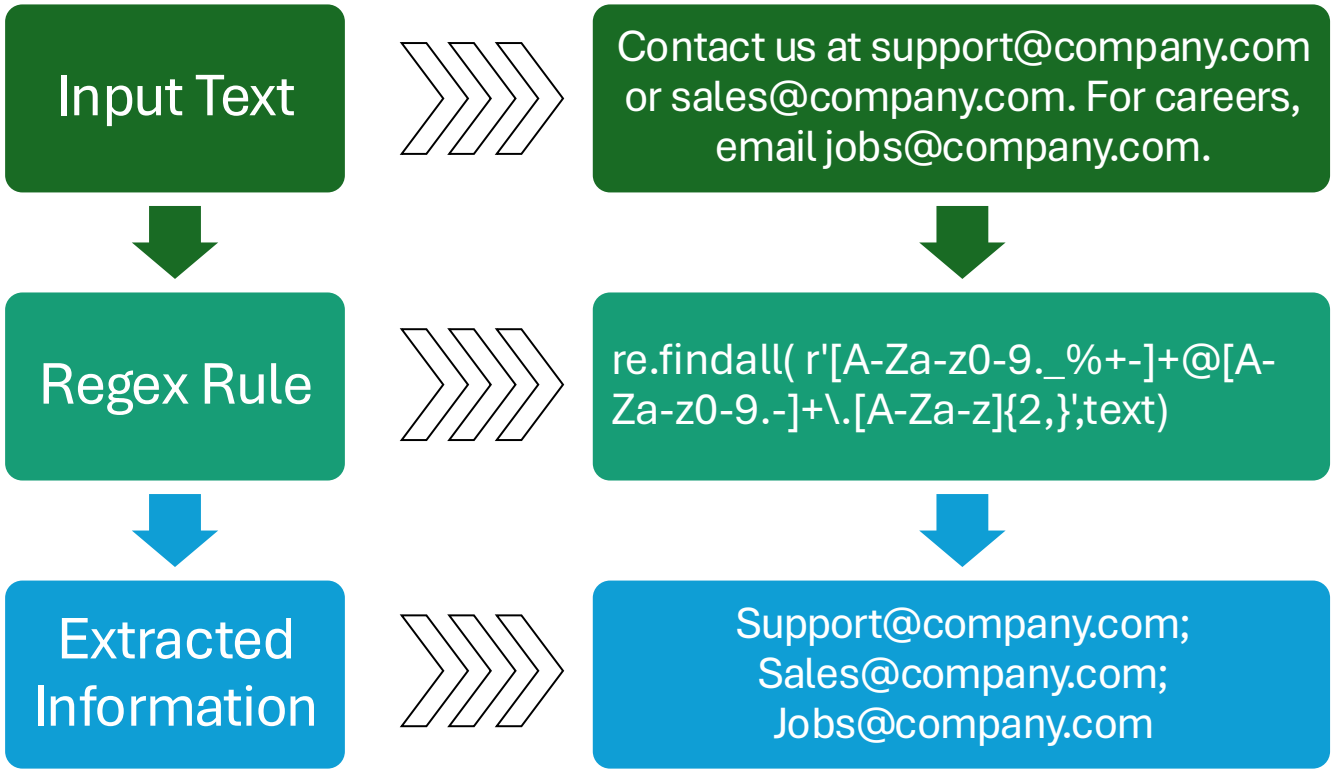
Key Functions and Parameters

- **re.search(pattern, string, flags=0)**
- **re.findall(pattern, string, flags=0)**
- **re.finditer(pattern, string, flags=0)**
- **re.sub(pattern, repl, string, count=0, flags=0)**
- **re.split(pattern, string, maxsplit=0, flags=0)**
- **re.compile(pattern, flags=0)**

Common Flags

•re.IGNORECASE (or re.I): Case-insensitive matching
•re.MULTILINE (or re.M): ^ and $ match at line boundaries
•re.DOTALL (or re.S): . matches newlines
•re.VERBOSE (or re.X): Allow comments and whitespace in patterns

Example: re.findall()



Other rule-based methods

Semantic / Logic-Based Rules: Specify how meaning is interpreted or inferred based on predefined logical or domain-specific rules

Context-Free Grammars (CFG): define hierarchical syntactic rules that describe how valid sentences or structures are formed

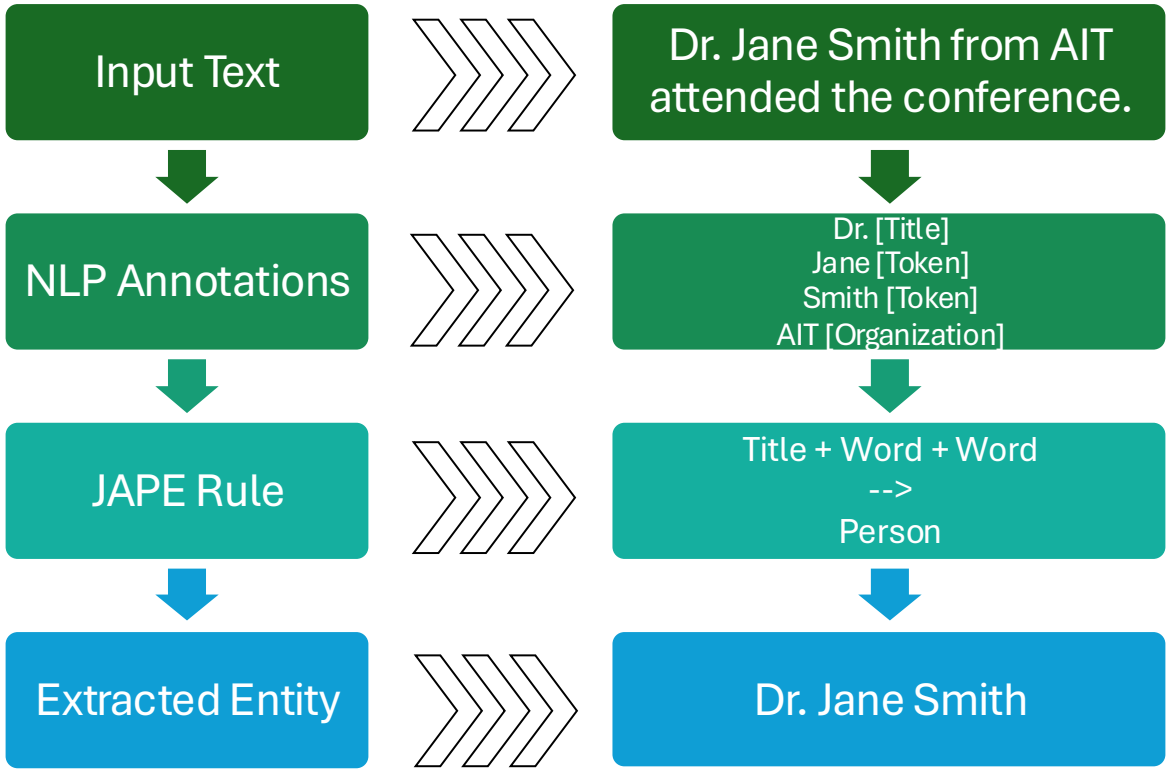
Finite State Machines (FSM): mathematical models of computation that can be used to recognize patterns in text

Other rule-based methods (cont.)

Tools

- **Regular Expression Engines**
- **NLP Libraries**
- **Text Extraction Frameworks**
- **GATE (General Architecture for Text Engineering)**
- **UIMA (Unstructured Information Management Architecture)**

Example: GATE JAPE Name extraction



Methods, Models, and Practice

- Classical machine learning
- Sequence Learner
- Transformer Models

Classical machine learning methods

Supervised learning

Models are trained using labeled data

Very effective in:

- Named Entity Recognition (NER)
- Text Classification
- Sentiment Analysis
- Relation Extraction

Unsupervised learning

Model tries to find hidden patterns or intrinsic structures using unlabeled data.

Very effective in:

- Clustering
- Topic Modeling
- Anomaly Detection
- Relation Extraction

Classical machine learning methods (cont.)

Common Models

- Naïve Bayes
- Support Vector Machines (SVM)
- Conditional Random Fields (CRF)
- Maximum-Entropy Markov Models (MEMMS)

Advantages

- Less manual rule writing
- Learns from data
- Often more adaptable than purely rule-based systems

Disadvantages

- Requires labeled training data
- Performance depends on feature engineering
- Can struggle to generalize across domains

Sequence learners

Acronyms	Full Name
CRF	Conditional Random Fields
HMM	Hidden Markov Models
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory
BiLSTM-CRF	Bidirectional LSTM-CRF

Sequence learners process text as a sequence and predict labels for each word based on its surrounding context

Sequence learners (cont.)

Example input text:

"Jenna Johnson worked at AIT in 2015."

Advantages:

- Captures contextual information and sequential dependencies
- Effective for NER and IE
- Learns patterns automatically from training data

Disadvantages:

- Requires labelled training data
- Can be computationally expensive to train
- May struggle with long-range dependencies

Word	Label
Jenna	PERSON
Johnson	PERSON
worked	O
at	O
AIT	ORGANISATION
in	O
2015	DATE

Transformer models

Deep learning models that use attention mechanisms to process an entire sentence at once instead of reading words sequentially

Text is first converted into numerical representations called embeddings (vector representations of words or subwords)

The encoder creates contextual representations for each token in the input sequence

The decoder generates an output sequence token by token using these contextual representations.

Transformer models (cont.)

State-of-the-art models:

- BERT
- GPT
- T5 (Text-To-Text Transfer Transformer)

Strengths:

- Cornerstone of modern text analytics due to their ability to handle long-range dependencies and complex relationships with text
- Excels in capturing intricate patterns and relationships within large-scale data

Weaknesses:

- Expensive to train
- Consume large memory
- Need large datasets
- Have input length limits, making very long text handling difficult
- Risk of overfitting if not properly regularized

Methods, Models, and Practice

- Modern approaches using general purpose models and Retrieval-Augmented Generation (RAG)

Modern General Purpose Models (LLMs)

Examples:

GPT-4

Claude

Gemini

Advantages

- Zero-shot and few-shot learning
- No feature engineering
- Can perform many NLP tasks with the same model
- Flexible across domains

Disadvantages

- May hallucinate information
- Less predictable than specialized systems
- Computationally expensive

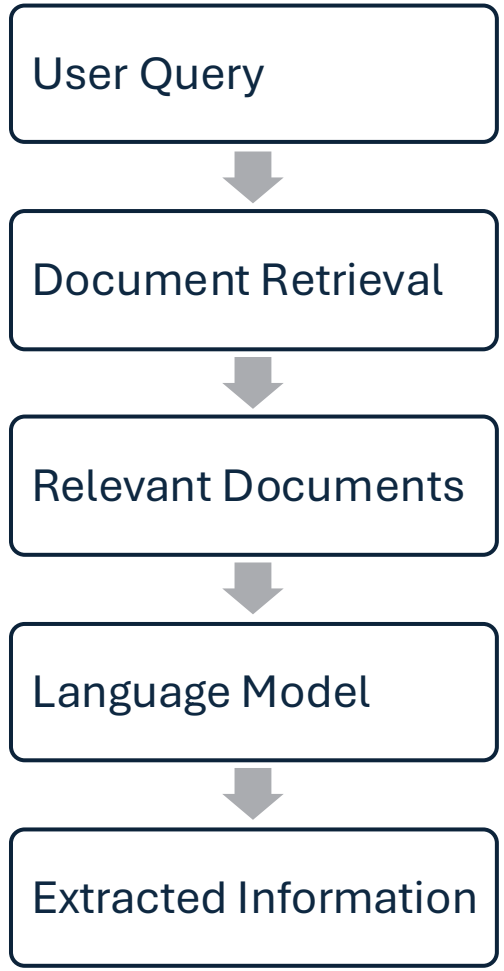
Retrieval-Augmented Generation

Advantages

- Uses external and up-to-date information
- Reduces hallucinations
- Works with large document collections

Disadvantages

- Retrieval quality affects results
- More complex architecture
- Higher latency than standalone models



Building Knowledge from Text

Methods, Models, and Practice

Heidi Hobel

AIT Austrian Institute of Technology | Research Engineer

AI:AT Factory | AI & Data Services Expert

Natural language processing & machine learning

Semantic search

Information extraction

Knowledge-based decision support

Transforming unstructured text into structured knowledge

Annotation, classification, and entity & relation extraction

Funded by

under discussion with

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).